Année 2020-2021
Ecole des Mines de Nancy
Tronc commun 2A

Nom, Prénom :

Test Analyse de Données

5 janvier 2021 - Durée : 2h
Cours et TD/TP autorisés
Sujet de S.Ferrigno (séances 1 à 6)
Durée conseillée sur cette partie : 1h30

Consignes : Veuillez répondre directement aux questions sur la feuille d'énoncé dans les espaces laissés libres à cet effet. Toutes les réponses devront être justifiées.

En physiologie, un moyen pour étudier la forme physique est de savoir à quelle vitesse le corps peut absorber et utiliser l'oxygène. Des sujets ont participé à un exercice prédéterminé qui consiste en une course à pied de 2km400. On a alors enregistré des mesures de leur consommation d'oxygène ainsi que d'autres variables continues telles que leur âge, leur pouls et leur poids. Des chercheurs se sont intéressés à comment pourrait on prédire la consommation d'oxygène à partir de ces variables. **fitness** (issu de Rawlings (1998)) est le jeu de données qui contient ces informations, en particulier les variables suivantes :

Name : Nom du participant.
Gender : Sexe du participant.
Runtime : Temps en minutes pour parcourir 2km400.
Age : Age (en années) du participant.
Weight : Poids (en kg) du participant.
Oxygen_Consumption : Mesure de la capacité à utiliser l'oxygène qui se trouve dans le sang.
Run_Pulse : Pouls à la fin de la course.
Rest_Pulse : Pouls au repos.
Maximum_Pulse : Pouls maximum durant la course.
Performance : Note globale de la forme physique.

Ces données ont déjà été utilisées lors du TD/TP 2 sur la Régression linéaire. Vous trouverez donc le jeu de données sur Arche, dans la partie Travaux pratiques, dossier TP2. Pour importer ces données sous R, taper la commande suivante :

$$fitness <- read.table("fitness.txt", header = T)$$

En mettant en oeuvre au mieux vos connaissances acquises en cours et en TD/TP et avec l'aide du logiciel R, répondez aux questions suivantes.

Partie 1 (10 points) : Etude descriptive et Analyse en composantes principales (ACP).

0.5pt 1. Donner les quartiles de la variable Oxygen_Consumption.

$Q_1 = 44.97$
$Q_2 = 46.77$
$Q_3 = 50.43$

1 pt 2. Existe-t-il un lien entre les variables Runtime et Age ? Pour répondre à cette question, vous utiliserez un test statistique et vous donnerez les hypothèses du test, la statistique de test et la valeur de la pvalue. Vous conclurez en utilisant un risque de 5%.

$H_0 : \rho = 0$
$H_1 : \rho \neq 0$

Statistique de test : 1.072

pvalue = 0.2926 > 5% donc non rejet de H_0.

1 pt 3. Reprendre la question 2. pour les variables Performance et Rest_Pulse.

$H_0 : \rho = 0$
$H_1 : \rho \neq 0$

Statistique de test : -2.9431

pvalue = 0.006335 donc rejet de H_0.

0.5pt 4. Quel serait l'impact des conclusions émises dans les questions 2. et 3. sur l'écriture d'un modèle de régression linéaire visant à expliquer la variable Oxygen_Consumption en fonction des autres variables quantitatives du jeu de données ? Il n'est pas nécessaire d'écrire ce modèle.

× Performance et Rest_Pulse étant très corrélées, elles pourraient ne pas intervenir ensembles dans le modèle

× En revanche, Runtime et Age n'ayant pas de lien, elles pourraient apparaître ensembles dans le modèle.

Lancer l'Analyse en composantes principales normée sur ce jeu de données en prenant soin de ne conserver que les variables qui se prêtent à ce type d'étude (c'est à dire en sélectionnant toutes les variables sauf Name et Gender).

0.5pt 5. Quels sont les individus de cette étude d'ACP ? Dans quel espace sont-ils définis au départ ?

Les individus sont les personnes ayant participé à la course à pied - Ils sont définis dans $\mathbb{R}^8$

0.5pt 6. Une ACP **normée** était-elle nécessaire dans le contexte de cette étude ? Justifier votre réponse.

Oui car les variables ont des unités différentes

1.5pt 7. Pour cette étude, combien d'axes allons-nous conserver pour effectuer la projection des données de départ selon le critère de Kaiser ? selon la règle du coude ? selon votre "bon sens" ? Justifier vos réponses.

- Kaiser : 2 axes car 2 valeurs propres > 1 (la 3ème valeur propre est tout de même proche de 1)
- Règle du coude : 3 axes car un "coude" apparait sur le graphe concerné à ce niveau
- Bon sens : 3 axes car on récupère 82% de l'information de départ

Dans la suite, nous travaillerons dans le plan $(1,2)$.

1 pt 8. Quelles sont les variables qui contribuent le plus à la construction de l'axe 1 ? de l'axe 2 ? Justifier vos réponses.

Axe 1 Runtime ≈ 22.23
Oxygen_Consumption ≈ 21.26
Performance ≈ 22.43

Axe 2 Age ≈ 31.30
Maximum_Pulse ≈ 25.84457

On regarde les contributions !

1 pt 9. Quelles sont les variables qui sont bien représentées dans le plan $(1,2)$? Justifier votre réponse.

Toutes sauf Rest_Pulse et Weight (voir le graphe de l'hypersphère unité dans ce plan)

0.5 pt 10. Parmi les 31 personnes qui ont participé à l'étude, quelle est la mieux représentée dans le plan $(1,2)$? La moins bien représentée ? Justifier vos réponses.

× Naomi est la mieux représentée avec un cosinus carré de 0.95 à peu près (4ème obs)

× Ralph est le moins bien représenté avec un cosinus carré de 0.014 à peu près (19ème obs)

1 pt 11. Quelle interprétation de l'ensemble des variables peut-on avoir dans le plan (1, 2) ?

× L'axe 1 oppose Performance et Oxygen-Consumption à Runtime. Performance et Runtime sont les 2 variables qui expliquent le mieux Oxygen-Consumption. Plus on court vite, plus on est performant physiquement.

× L'axe 2 oppose Age à Maximum-Pulse. En endurance, plus on est "agé", plus on a un pouls maximum qui "baisse" en course.

Nous travaillons à présent dans le plan (1, 3).

0.5 pt 12. Quelle est la variable qui contribue le plus à la construction de l'axe 3 ? Justifier votre réponse.

Il s'agit de la variable Weight avec une contribution de 82.99.

0.5 pt 13. Donner les coordonnées de l'individu 26 dans le plan (1, 3). Cette personne est-elle bien représentée dans ce plan ? Justifier vos réponses.

Coordonnées sur l'axe 1 : 1.389

l'axe 3 : -0.215464

Cosinus carré dans le plan (1, 3) ≈ 0.9442 qui est proche de 1 donc l'individu est bien représenté dans le plan.

Partie 2 (6 points) : Classification Automatique Hiérarchique (CAH).

Lancer à présent une classification hiérarchique ascendante sur le jeu de données fitness en ne prenant pas en compte, comme pour la Partie 1, les variables Name et Gender et en utilisant la méthode de Ward. Dans cette partie, nous travaillerons avec la distance Euclidienne qui est la distance utilisée par défaut dans les diverses fonctions R dont vous aurez à vous servir.

0.5pt 1. Quelle est la distance entre les individus 21 et 31 ?

5.9426302 (via matrice des distances)

0.5pt 2. Donner le code R qui vous a permis d'effectuer cette classification à partir de la matrice des distances.

```
cah.ward <- hclust(fitness.d, method="ward.D2")
```

fitness.d : matrice des distances

0.5pt 3. Combien de classes choisiriez-vous à partir du dendrogramme ?

3 ou 4 classes

Dans la suite nous travaillerons avec 4 classes.

1pt 4. Parmi les 31 personnes de l'étude, combien sont affiliés à la classe 1, la classe 2 ? Quels sont les individus qui font partie de la classe 4 ?

Classe 1 a 4 individus
Classe 2 a 12 individus
Classe 4 a individus 24, 27, 30, 31

0.5pt 5. A quelle classe appartient l'individu numéro 12 ?

Classe 2

1 pt 6. Quelle est la moyenne et l'écart-type de la variable Performance dans chacune des quatre classes ?

	Classe 1	Classe 2	Classe 3	Classe 4
Moyenne	13	8.666...	7.181818	3.25
Ecart-type	0.8164966	1.154701	2.227922	2.753785

Lancer à présent une classification hiérarchique ascendante sur l'ensemble des composantes principales obtenues en effectuant l'ACP sur le jeu de données bfitness (sans les variables Name et Gender).

1 pt 7. Combien de classes suggère le dendrogramme ? Justifier votre réponse.

3 classes (voir graphe du gain d'inertie)

1 pt 8. Quel est l'individu le plus représentatif de la classe 1. Justifier votre réponse.

L'individu n° 2 car sa distance au centre de gravité de la classe est la plus faible (1.314889)

Partie 3 (4 points) : Analyse discriminante. 1 point par question

Dans cette partie, nous travaillerons avec le jeu de données fitness de départ (mais sans la variable Name). L'idée est d'utiliser les 8 variables quantitatives du jeu de données pour retrouver le genre des personnes qui ont participé à cette étude (variable Gender).

1. Classer les 8 variables quantitatives de la plus discriminante à la moins discriminante. Justifier votre réponse.

Weight pvalue = 0.000779

Oxygen-Consumption pvalue = 0.00575

Runtime pvalue = 0.0154

Performance pvalue = 0.0185

Rest-Pulse pvalue = 0.107

Run-Pulse pvalue = 0.184

Maximum-Pulse pvalue = 0.24

Age pvalue = 0.747

2. Réaliser un test de Wilks sur ces données. Vous préciserez les hypothèses du test, la valeur de la statistique de test, la pvalue et vous conclurez en prenant un risque de 5%.

$H_0 : \mu_1 = \mu_2$ (avec μ_j = moyenne de la densité de proba du groupe j)

$H_1 : \mu_1 \neq \mu_2$

Statistique de test = 0.48819

approximée par $F = 2.8831 \sim F(8, 22)$

pvalue = 0.02337 < 5% donc rejet de H_0.

3. Réaliser une analyse discriminante avec l'ensemble des variables du jeu de données. Combien d'axes discriminants sont retenus ? Si on applique ce modèle aux données de départ, quel est la taux d'erreur ?

1 seul axe !

Erreur ≈ 38.7%

4. Prédire à quel genre (et avec quelle probabilité) appartient une personne ayant les caractéristiques suivantes (10.06, 45, 75.5, 50.55, 178, 65, 185, 8) dans l'ordre d'apparition des variables dans le jeu de données initial ?

Féminin avec la probabilité ≈ 60%

(on utilise le modèle de la question précédente)